COMPARISON BETWEEN HALFWAY REALISTIC-LOOKING PHYSICAL MODELS OF HUMAN VOCAL TRACT

Takayuki Arai

Sophia University (Japan)
arai@sophia.ac.jp

ABSTRACT

Two versions of the physical model of the human vocal tract have recently been developed. They are halfway realistic-looking and have lips, teeth, a tongue, velum, pharyngeal wall, etc. so that users can learn their positions during articulation. Because parts of the side and rear plates are transparent, the inside of the vocal tract is visible. Furthermore, some details of the anatomy have been simplified or ignored so that users can focus on the important aspects of speech production. One of the versions is static and produces the vowel /a/. The other version has a flexible tongue for changing the configuration of the vocal tract. In this study, we compared the two versions including their acoustic outcomes. We confirmed that 1) both models can produce clear /a/ and 2) more vowels can be produced with the model with a flexible tongue.

Keywords: vocal-tract models, vowel production, acoustic characteristics, education in phonetics.

1. INTRODUCTION

Different types of the physical models of the human vocal tract have been developed [1–3]. The main purpose of these models is to help learners of phonetics and speech science understand the mechanisms of speech production. However, these models are now applied for other purposes, such as, basic research, language learning, speech pathology, and clinical applications. Different models are currently used depending on individual needs.

Table 1: Grouping of previously developed physical models of human vocal tract

	Straight	Bent
Static	e.g., VTM-T20 	e.g., Head-shaped
Dynamic	e.g., VT M-S20 	e.g., Flexible-tongue

Table 1 shows such vocal-tract models grouped based on two dimensions; straight vs. bent and static vs. dynamic. Because the human vocal tract is bent and dynamic, a flexible-tongue model is more realistic. However, when we teach how the static shape of a vocal-tract configuration determines its frequency characteristics, static models, such as the head-shaped model, are useful. Furthermore, we sometimes want to focus on the relation between a simplified cross-sectional area function and sound quality; therefore, static and straight models, such as VTM-T20, are suitable. On the other hand, if we want to focus on the dynamic aspects of speech production, straight and dynamic models, such as VTM-S20, are suitable.

Two bent vocal-tract models have recently been developed [4, 5], which were inspired by anatomical models for medical purposes. They are anatomical because they have lips, teeth, a tongue, velum, pharyngeal wall, etc. In that sense, users, including students who are studying phonetics or speech science, can learn how these speech organs are placed. Parts of the side and rear walls are transparent, so that the inside of the vocal tract is more visible. In many anatomical models, some details of the anatomy are ignored or simplified, enabling users to focus on the important aspects of the organs. One of the two models is a static model, with which the vocal-tract configuration cannot be changed; it is set to for one vowel, i.e., /a/ [4] (hereafter, 2017 model). The other model has a flexible tongue; therefore, we are able to change the configuration of the vocal tract [5] (hereafter, 2018 model). In this study, we compared these two halfway realistic-looking models since the details of these models have not been investigated. Since they were designed to produce speech sounds, including vowels, we particularly compared their acoustic outcomes and confirmed that both models can clearly produce /a/ and the 2018 model can produce several more vowels.

Figure 1: 2017 model [4]. (a) Overview, (b) views from different angles.

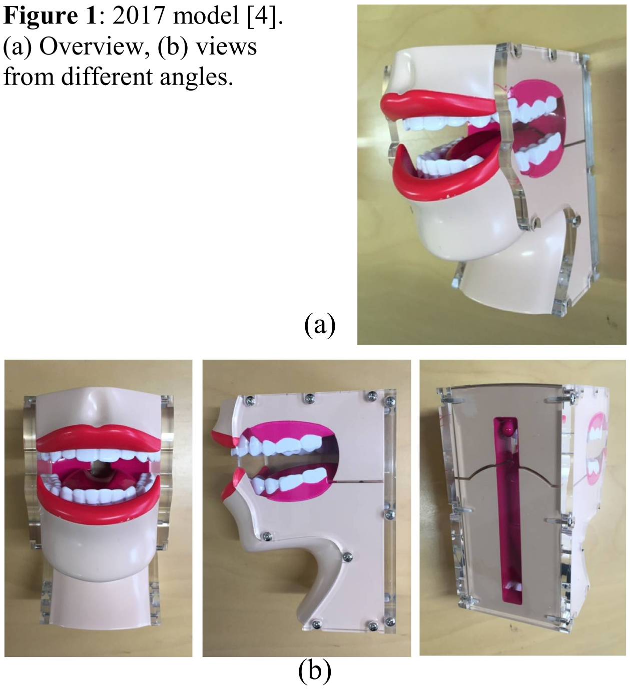


Figure 2: 2018 model [5]. (a) Overview, (b) views from different angles.

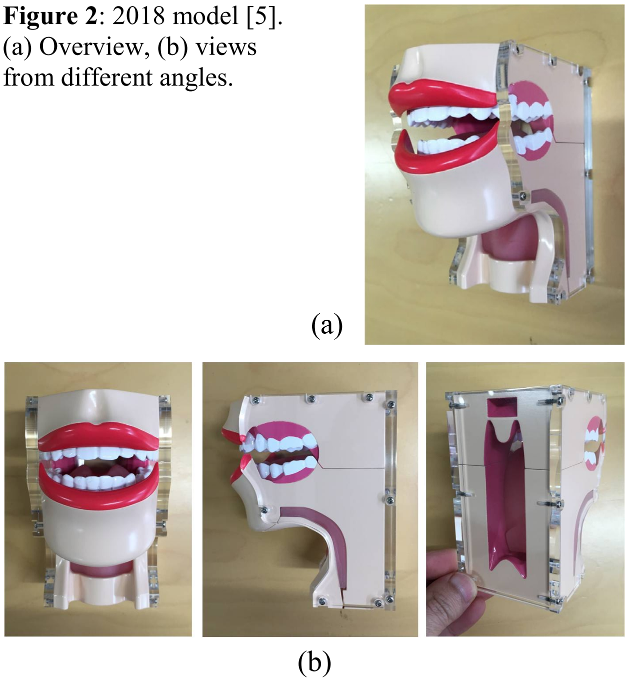
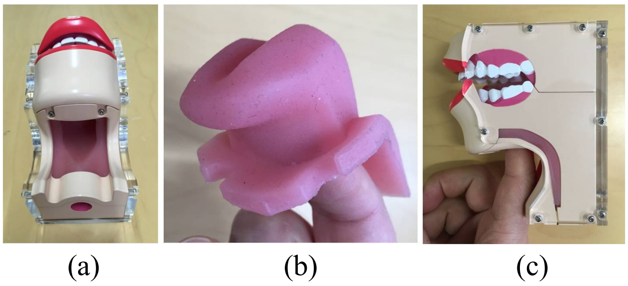


Figure 3: 2018 Model [5]. (a) View from low angle. (b) Tongue placed on index finger. (c) One way to manipulate tongue.



2. 2017 MODEL

Figure 1 shows the 2017 model [4]. It is static and the position of each speech organ is based on the vocal-tract configuration for producing /a/. As shown in this figure, one can easily recognize the lips, teeth, and tongue, and the mouth is wide open. The cheeks are not included because the width of the model without the side plates is set to 50 mm, which is realistic to cover the major vocal tract. Instead of cheeks, transparent acrylic side plates cover the side sections of the vocal tract. The plates are transparent so that the inside of the oral cavity, including the tongue, is visible from the outside. The rear plate for the anterior pharyngeal wall is also made of transparent acrylic, and the uvula and tongue root are also visible from the back of the model. The laryngeal end of the vocal tract has a hole, which can be connected to a sound source to produce /a/. The materials, except the above-mentioned acrylic plates, were made using a 3D printer, AGILISTA, and the surface was painted to differentiate the different parts of the model.

3. 2018 MODEL

Figure 2 shows the 2018 model. This version is similar to the 2017 model; however, it is a dynamic model, and the vocal-tract configuration can be varied by deforming the tongue. The tongue is made of a gel-type material, a polyethylene-styrene copolymer, with an ASKER-C hardness of 2 and 4. The other parts of the model were made of the same materials those of as the 2017 model.

Figure 3 shows the tongue and how it is housed in the model. From the angle shown in Fig. 3(a), the tongue is reachable from the lower jaw. The tongue can be deformed by manipulating it from underneath the vocal tract. There is a hole at the back of the tongue, and one can insert a finger for high vs. low deformation, for example. The tongue has a thin and long semi-circular groove at the center, and this is important to produce high vowels, such as /i/. Because the tongue root is extended to the anterior pharyngeal wall, when one pushes the tongue root against the pharyngeal posterior wall, we can hear back vowels, such as /a/.

4. ACOUSTIC ANALYSIS

4.1. Recordings

We tested the output signals produced with the two models. A whistle-type artificial larynx was used as a sound source. Output signals from the models were digitally recorded using a microphone (Sony, ECM-23F5) and digital audio recorder (Marantz, PMD670). The original 48-kHz sampling frequency for the recordings was downsampled to 8 kHz for acoustic analysis.

For the 2017 model, a single configuration of the tongue was tested for the recordings. For the 2018 model, on the other hand, five different configurations of the tongue were tested to simulate five Japanese vowels.

4.2. Vowels /a/ with Two Models

Figure 4(a) shows a spectrum calculated with the 30-ms Hamming window of an output signal produced with the 2017 model with a whistle-type artificial larynx. This spectrum clearly shows the frequency characteristics of /a/, such as the first formant (F1) frequency of approximately 850 Hz and second formant (F2) frequency of approximately 1250 Hz.

Figure 4(b), on the other hand, shows a spectrum calculated from an output signal produced by the 2018 model with an /a/ configuration, where the tongue root was placed almost at the posterior wall. This spectrum also has the frequency characteristics of /a/, as shown in Fig. 4(a); the F1 frequency is approximately 900 Hz and F2 frequency is approximately 1250 Hz. However, the F2 amplitude is a little bit lower in Fig. 4(b) compared to that in Fig. 4(a). The possible reasons for this difference might be due to the following points:

- length of the oral cavity.
- size of the lip aperture.
- shape of the tongue.

Figure 4: Spectra of output signals produced with 2017 and 2018 models with /a/ configuration.

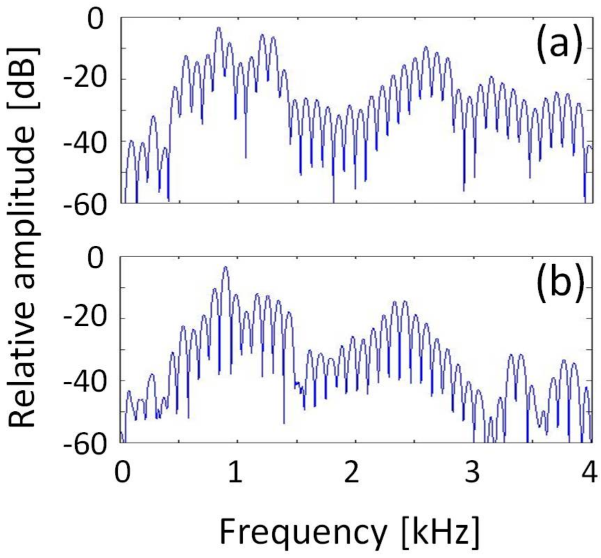


Figure 5: Mid-sagittal cross-sectional views of 2017 model (a) for /a/ and 2018 model (b) with default setting.

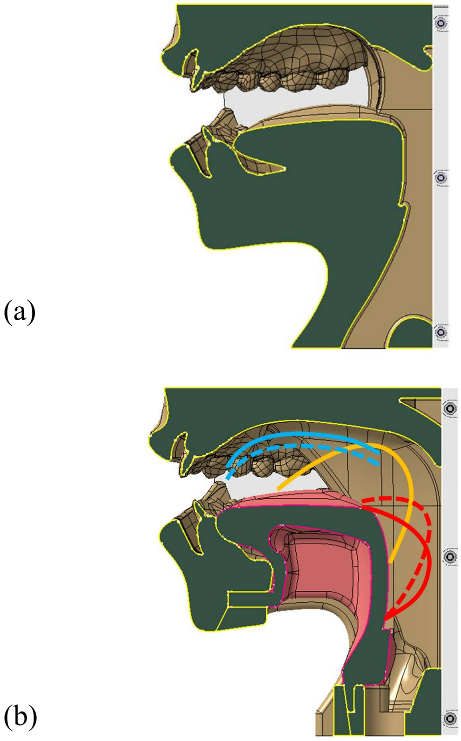
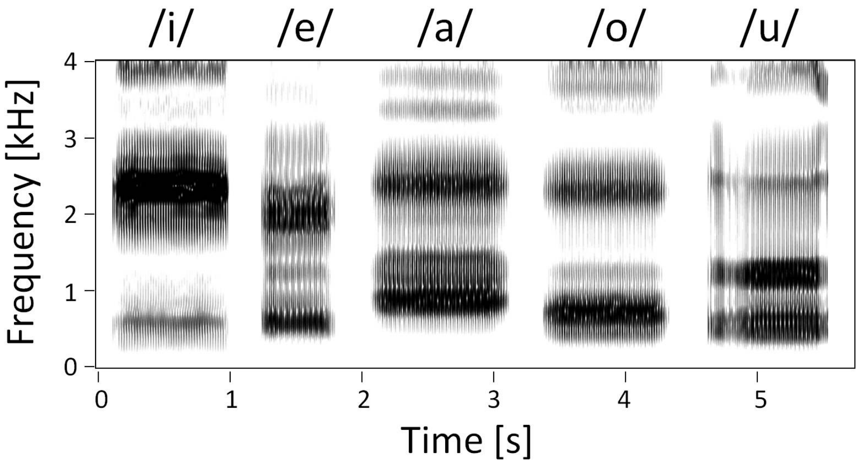


Figure 5 shows the mid-sagittal cross-sectional views of the 2017 model (a) and 2018 model (b). From this figure, we can discuss the length of the oral cavity (from the pharyngeal wall to the upper incisor). This length in the 2017 model is approximately 77 mm, while it is approximately 73 mm in the 2018 model. The size of the lip aperture of the 2017 model is 24 mm, while it is approximately 19 mm in the 2018 model. However, the most dominant contribution should be the shape of the tongue because the one in the 2018 model can be varied. One of the typical configuration of the tongue is shown as the red solid curve in Fig. 5(b). Thus, the acoustic outcome moderately changes depending on its shape.

4.3. Other Vowels Produced with 2018 Model

Figure 6 shows a spectrogram of output signals produced with the 2018 model with a whistle-type artificial larynx. Vowel /a/ in the middle of this figure is from the same utterance analyzed in Section 3.2 (the duration of the utterance was shortened). For /i/, the tongue was heavily raised in the oral cavity, as the blue solid curve in Fig. 5(b). When the level of this tongue was decreased, /e/ was produced (the blue dashed curve in the same figure). For /o/ and /u/, we reduced the lip aperture by placing fingers or putting clay to the sides of the mouth opening. With this reduced lip aperture, the

Figure 6: Spectrogram of output signals produced with 2018 model with different tongue configurations.



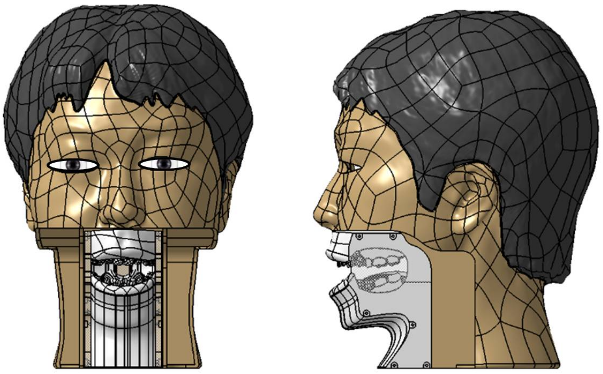
tongue root was pushed back against the pharyngeal wall for /o/, as was done for /a/. The position of the tongue root was, however, approximately 10 mm higher for /o/ (the red dashed curve in Fig. 5b) than for /a/ (the height was approximately 40 mm above the larynx for /a/). For /u/, the middle of the tongue was diagonally raised against the corner of the bent vocal tract (the orange solid curve in Fig. 5b). From these spectra and by listening to the output sounds, this model produced acceptable vowel qualities.

5. DISCUSSION AND CONCLUSION

In the previous section of acoustic analysis, we have seen that the 2017 model can clearly produce /a/ and the 2018 model can produce different vowels depending on the vocal-tract configuration. Such vocal-tract models should be widely used in many applications, such as education in phonetics and/or speech science, language learning, and clinical settings in speech-language pathology. However, to the best of our knowledge, there are no similar models in terms of both design and sound quality.

Hofe (2011) conducted biomimetic vocal-tract modeling for his dissertation [6]. His model can produce speech sounds from the anatomical structure. Similar but more engineering-oriented speaking robots have been developed [7, 8]. The goal of the present study was, however, for education or clinical purposes, and it is crucial that such a model be compact enough and can be manipulated in addition to being realistic-looking. Furthermore, the output sounds should be intelligible so that users can easily recognize the difference in sounds and configurations of the vocal tract at the same time. In this sense, the two previously proposed models meet these criteria. Especially, the 2018 model can also be applied to produce consonantal sounds.

Figure 7: Designed of a head model with 2017 model.



The 2018 model has also potentials for a basic research purpose. For example, the issue of "the tongue bracing" is one of them. It is reported that the tongue is braced against the teeth and/or the palate during speech production [9, 10]. With the 2018 model, we can test that the tongue is actually braced when vowels /i/ and /e/, for instance.

In the future, more models should be designed along this concept, such as one with a movable jaw. In addition, the head model is currently designed, so that users can combined it with the 2017/2018 models to complete a more realistic-looking as shown in Fig. 7.

6. ACKNOWLEDGMENTS

This work was partially supported by JSPS KAKENHI Grant Number 18K02988.

🔊 **2017 Model** 🔊 **2018 Model**

7. REFERENCES

[1] Arai, T. 2007. Education system in acoustics of speech production using physical models of the human vocal tract. *Acoust. Sci. & Tech.* 28, 190–201.

[2] Arai, T. 2012. Education in acoustics and speech science using vocal-tract models. *J. Acoust. Soc. Am.* 131, 2444–2454.

[3] Arai, T. 2016. Vocal-tract models and their applications in education for intuitive understanding of speech production. *Acoust. Sci. & Tech.* 37, 148–156.

[4] Arai, T. 2017. Vocal-tract model with stati articulators: Lips, teeth, tongue, and more. *Proc. of INTERSPEECH*, Stockholm, 4028–4029.

[5] Arai, T. 2018. Flexible tongue housed in a static model of the vocal tract with jaws, lips and teeth. *Proc. of INTERSPEECH*, Hyderabad, 171–172.

[6] Hofe, R. 2011. *Biomimetic Vocal Tract Modelling: An Artificial Speaker to Investigate the Energetics of Speech Production.* PhD Thesis, the University of Sheffield.

[7] Fukui, K., Kusano, T., Mukaeda, Y., Suzuki, Y., Takanishi, A. and Honda, M. 2010. Speech robot mimicking human articulatory motion. *Proc. of INTERSPEECH*, Makuhari, 1021–1024.

[8] Brady, M. C. 2010. Prosodic timing analysis for articulatory re-synthesis using a bank of resonators with an adaptive oscillator. *Proc. of INTERSPEECH*, Makuhari, 1029–1032.

[9] Stone, M. 1990. A three-dimensional model of tongue movement based on ultrasound and X-ray microbeam data. *J. Acoust. Soc. Am.* 87, 2207-2218.

[10] Gick, B., Allen B., Stavness, I. and Wilson, I. 2013. Speaking tongues are always braced. *J. Acoust. Soc. Am.* 134, 4204.